DOCUMENT RESUME

ED 356 266 TM 019 697

AUTHOR Chelimsky, Eleanor
TITLE Student Achievement Standards and Testing. Testimony before the Subcommittee on Elementary, Secondary, and Vocational Education of the Committee on Education and Labor, House of Representatives.
INSTITUTION General Accounting Office, Washington, DC. Program Evaluation and Methodology Div.
REPORT NO GAO/T-PEMD-93-1
PUB DATE 18 Feb 93
NOTE 15p.
PUB TYPE Legal/Legislative/Regulatory Materials (090) -- Reports - Evaluative/Feasibility (142) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Academic Standards; Achievement Tests; *Cost Estimates; Cross Cultural Studies; Educational Assessment; Elementary Secondary Education; Foreign Countries; National Competency Tests; *National Programs; Program Costs; Standardized Tests; *Student Evaluation; *Testing Programs; Test Use
IDENTIFIERS Alternative Assessment; Canada; *Performance Based Evaluation; Standard Setting; United States

ABSTRACT

The General Accounting Office (GAO), at the request of the House Committee on Education and Labor, Subcommittee on Elementary, Secondary, and Vocational Education, conducted studies on the extent and cost of testing in the United States, the experience of Canada in testing, and initial efforts to set standards for judging student performance on the National Assessment of Educational Progress (NAEP). Main findings and conclusions from the first two studies are presented. A survey of all states and a national sample of school districts has suggested that U.S. students do not seem to be overtested, spending about seven hours a year in testing (including preparation and related activities). Nationally, systemwide testing in 1990-91 cost about $516 million. A national performance test could be expected to cost about $330 million. Regional state clusters of performance tests, as recommended by the National Council on Educational Standards and Testing would add about $193 million to current costs, with 25 minutes additional testing time. The Canadian testing system features a coordinated set of standards, course specifications, and tests that are well regarded by both educators and the public. In reviewing general questions of testing policy, the importance of teacher and administrator involvement is emphasized. Another major issue is that of ensuring the technical quality of any tests in a national system. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

United States General Accounting Office

GAO

Testimony

Before the Subcommittee on Elementary,
Secondary, and Vocational Education
Committee on Education and Labor
House of Representatives

For Release on Delivery
Expected at 10:00 a.m.
Thursday
February 18, 1993

Student Achievement Standards and Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED356266

Statement of Eleanor Chelimsky
Assistant Comptroller General
Program Evaluation and Methodology Division

TM019697

BEST COPY AVAILABLE

Mr. Chairman and Members of the Committee:

I am pleased to be here today to discuss GAO's work in the broad area of student achievement standards and testing. At your request, we have done three studies: one on the extent and cost of testing in this country, another on the experience with standards and tests in Canada, and a third on the initial efforts to set standards for judging student performance on the National Assessment of Educational Progress (NAEP). A report on the first is available and a report on the second will be published soon.[1] We issued an interim paper on the NAEP work last March.[2] We expect the final report to be issued in 90 days.

I will focus today on the main themes of our findings and conclusions from the first two studies. Our reports describe the scope of the work and methods we used in detail. In brief, we gathered data on the present extent and costs of testing in the United States (and the views of education officials on testing issues) by surveying all the states and a national sample of school districts. We also estimated likely costs for a national test. With regard to the Canadian experience with standards and tests, our effort involved reviewing provincial evaluations and other data, visiting provincial and district offices in several provinces, and interviewing officials in the provinces that we could not visit. The Canadian experience is relevant to the current U.S. effort to establish standards and related tests for school learning because some provinces have for some time had testing systems similar in various ways to plans suggested for the United States and because standards play a large role in those systems.

I will turn first to the information we produced on current testing and our forecasts of resources required for a national test and then discuss the Canadian experience.

TESTING TODAY

In 1990-91, students in the United States did not seem to have been overtested--the average student spent only 7 hours annually on systemwide testing (including preparation, test-taking, and all related activities)--and the cost totaled, on the

---

[1]U.S. General Accounting Office, Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination, GAO/PEMD-93-8 (Washington, D.C.: January 1993), and Educational Testing: The Canadian Experience With Standards, Examinations, and Assessments, GAO/PEMD-93-11 (Washington, D.C.: February 1993).

[2]U.S. General Accounting Office, National Assessment Technical Quality, GAO/PEMD-92-22R (Washington, D.C.: March 1992).

average, $15 per student, including the cost of the test and staff time.[3] The bulk of this testing was traditional in format (71 percent of tests consisted of multiple-choice questions only). Newer test types, such as performance tests in which students write out some answers, were much less common: tests with more than just a writing sample element were in use in only seven states. The performance tests also cost more. In the states where we had the best comparative data we found that multiple choice tests averaged less than half the cost of performance tests--$16 versus $33 per student, respectively. We estimated that in the nation as a whole, systemwide testing in 1990-91 cost about $516 million.

ESTIMATES OF NATIONAL TESTING OPTIONS' COSTS

We used our data on current costs for different kinds of tests to estimate what it would cost for a national-level test (assuming three grades tested in a year, totaling 10 million students). Since multiple-choice tests currently average about $16 per student, a national multiple-choice test would cost about $160 million. Because performance tests cost more (an average of $33 per student), national implementation of such a test, again at three grade levels, would cost a total of $330 million. Also, our data showed that these tests are expensive to develop: we estimate a national system would cost as much as another $100 million in one-time development costs.

The new costs of a national testing plan would vary, however, depending on whether schools added the test or used it to replace others currently in use. The multiple-choice option would add the least new cost in money and time, since (from data we gathered on past decisions) we predict three quarters of the districts would drop an existing test and replace it with the national test. Because many fewer districts use performance tests now, a national performance test would add more new costs in money and time: $209 million and another half hour per student per year. Regional state clusters of performance tests, the option recommended by the congressionally mandated National Council on Education Standards and Testing (NCEST), would add slightly less: $193 million of costs and 25 minutes more for the

[3]We define systemwide tests as those given to all, almost all, or a representative sample of students at any one grade level in a school district. This definition covers most standardized tests, except those given to certain groups. We did not include tests given to students under the requirements of the federal Chapter 1 program (unless districts gave such tests systemwide, which is common, according to Department of Education officials).

average student in testing time.[4]

TESTING OFFICIALS' VIEWS

Cost is not the only issue in comparing the options, of course. Multiple choice tests are familiar and provide strong comparative data but--according to opinion data from our survey--are least valued by state and local testing officials. State clusters of different performance tests are the least-developed method, cost twice as much as multiple choice tests, and would not necessarily be comparable among themselves or over time. They may, however, be better linked to local teaching and--again according to our survey--are viewed by testing officials as better measurements of what students know and can do.

Testing officials saw continued benefit to testing in general, even if there were to be more tests, but in discussing trends in the field, they expressed concern over the purpose, quality, and locus of control over further tests. With regard to purpose, our respondents voiced their preferences for more performance-based assessment that can help diagnose learning and teaching needs at the lowest levels, and they also recognized as valid the purpose of producing national data that are comparable over time. However, a third purpose of testing--accountability--was downplayed by our respondents, and concerns about possible misuses of tests in this regard (to compare unlike schools and districts, for example, or to reach unwarranted conclusions about students), were cited quite often as well. Quality and locus of control issues were expressed in respondents' preferences for tests that are of high technical quality, measure diverse skills in diverse ways, and cover what their teachers teach.

On the question of a national test or system of tests, our survey revealed significant opposition to the concept. Forty percent of local respondents and 29 percent of state respondents saw no advantages to a national system, and they forecast some disadvantages, particularly the potential for misuse of results. (Thirty-two percent of local respondents and 53 percent of state respondents did, however, specifically cite the potential for comparing test scores nationally as an advantage of a national testing system, although this purpose is to some degree in conflict with the local utility they also wanted.)

CONCLUSIONS

The costs of a national examination system may be less than anticipated. Assuming a hybrid system of testing for a number of

---

[4]The Council's recommendations are in its final report, Raising Standards for American Education (Washington, D.C.: January 1992).

potential purposes--testing all students in three grades--our estimates of the cost are higher than those of some national test proponents but lower than those of some opponents. Our projected figure of $330 million annually (for the most likely type of performance test, similar to tests in use in some states now) is about one tenth the amount some have suggested. The new costs would be less than that (about $200 million) and the added student testing time (increasing by up to 30 minutes the average amount of systemwide testing time per student, to a national average of about 7.5 hours total time and 4 hours of actual test-writing per student per year) does not seem unduly burdensome.

More specific forecasts or predictions will require making some decisions about the purpose or purposes that national tests can be expected to serve. Our data exemplify this need to choose in two ways. First, tension exists between our correspondents' preferences for two distinctly different emphases in testing: tests developed under local control and tests used principally for monitoring progress over time. Local control suggests a wide diversity of tests matched, in order to be most useful, to local variations in what is taught and learned; however, the goal of monitoring across classrooms, schools, districts, or states sets limits to the variation in tests that can be allowed without losing comparability. Second, tension exists between both local control and monitoring, on the one hand, and accountability, on the other. Although our respondents were not greatly concerned with accountability, others--chiefly outside the schools--have suggested that this purpose may be the most important: that is, using test results for high-stakes decisionmaking about students, teachers, or schools, and thereby emphasizing the importance of teaching and learning the material to be tested. Since it is not clear that one test can serve all three purposes, we conclude that decisions about test purposes are a high priority.

A final point is that the opposition we found to national testing, although abstract in the sense of not being linked to a particular proposal, should be carefully considered and addressed. The cooperation of state and local administrators and educators is important for any national testing effort. It seems reasonable to believe that if their knowledge, skills, and involvement can be effectively harnessed to the national testing effort, the success of the enterprise is more likely to be achievable.

EDUCATION STANDARDS AND TESTING IN CANADA

Turning to our second study for the committee, Mr. Chairman, let me present some observations drawn from the experience of Canadian provinces with education standards and testing, discussed in detail in our full report. As an affluent "high-tech" industrial society, Canada resembles the United States in many ways, and it also has considerable experience with a

decentralized student testing system presenting features recommended by NCEST for future adoption in the United States. Such features include measuring progress in relation to standards, using performance tests and other methods, and involving teachers intimately in all phases of testing. The United States does not lack experience with testing, of course, but what has happened in Canada affords useful contrasts on some key dimensions, as well as interesting information with regard to the development of incentive systems to counter various problems and pitfalls.

In brief, the major instructive contrasts and important elements we found are as follows.

Province-Level Standards

In Canada, educational standards are currently set at the province level, with major involvement of educators, especially teachers. (A recent effort there to set some national standards in basic learning areas, as a prelude to a national test of minimum competencies, has also included extensive involvement of teachers.) This differs from current efforts in the United States to set national standards chiefly by groups of experts, with only modest teacher involvement.

Different Tests for Different Purposes

In most Canadian provinces, two entirely different testing systems are dedicated to the separate purposes of certifying whether individual students meet standards (accountability) and tracking whether learning in general across a province is in line with what is expected (monitoring). (We refer to these in our report as examination and assessment systems, respectively; five provinces have the former and eight the latter.) This contrasts with the views of some in the United States who have proposed a single test or assessment method to serve many purposes.

Tests Linked to Standards

Both examinations (for accountability) and assessments (for monitoring) are developed within a province based on the standards for what should be learned in a particular course (in the case of the examinations) or in a particular subject and grade (in the case of the assessments). Both kinds of tests are revised often with major teacher involvement to reflect constant changes in those standards. This contrasts with the large U.S. use of commercially developed tests (customized in some cases to reflect state requirements) to measure students' cumulative knowledge of broad subject areas. In addition, both types of provincial tests use multiple methods, including the common use of essays but other tasks as well. The predominant format of U.S. tests, in contrast, is multiple-choice questions.

Stakes Differ for Different Tests

The idea often heard in recent U.S. testing debates that it is necessary to attach high stakes to all or many tests to emphasize the importance of learning to teachers and students does not seem to be reflected in the Canadian experience. For example, examination scores do not stand alone but are blended with teacher evaluations to form students' final grades, and the weight given to the exam score has been declining. (Assessments have no stakes for students or teachers.) Canada seems to rely instead on the continuous funded involvement of teachers in all phases of standard-setting and design of both examinations and assessments, as well as in test administration and scoring, to emphasize the importance of provincial standards.

Safeguards Associated With Tests

Canadian officials have employed a variety of safeguards to prevent misuse of test results. Safeguards in the examination system (where the accountability purpose means that results will have some consequences for students and teachers) include distributing the test specifications widely in advance, ensuring multiple opportunities for success, allowing for rescoring, and accommodating students with disabilities. In the assessment system--that is, tests designed to monitor or give an accurate picture of how all students are doing--other safeguards such as requirements that all students be tested and that reporting be both delayed and aggregated help ensure, on the one hand, data undistorted by biased participation and, on the other, fewer possibilities that results can be misused in decisions about individual students or teachers. Again, where data on all students are not needed in the assessments, sampling is increasingly used to permit multiple methods of testing (such as more expensive performance methods) without increases in cost.

Resources for Learning

Provincial funding formulas have been used in Canada to level resources among schools in a province and thus enable teachers generally to have comparable resources to implement the curriculum requirements. This is in contrast to sometimes large resource disparities among districts in the United States which give rise to the complaint that testing is inherently unfair since students may have experienced major differences in opportunity to learn. Thus, the issue discussed in the United States concerning "delivery standards," which some believe should accompany learning or achievement standards, is mitigated in Canada, because a degree of equalization of resources has been achieved.

Inadequate Evidence of Results

It is important to note that the effects of Canada's efforts to set standards and link tests to them have not been established. It is not known whether the elaborate strategy Canadian provinces have put in place has in fact caused better student achievement. No independent yardstick--no set of data, no national evaluation--affords such a measurement. There is some information on other effects of the effort, but it is scattered and of varying quality. For example, there are assertions by teachers that there has been some narrowing in what they teach and how, and there are survey data showing that high stakes on examinations elicit both anxiety and increased motivation in students. Increased fragmentation or stratification of student groups in some provinces has also been suggested to be the result of the isolation from others of those taking courses for which there are exams. Also, a rise in the number of students taking an extra year of high school is attributed, in part, to some staying longer in order to do better on the last set of examinations. In the view of some, and to the degree that they are accurate, all these results--from greater teacher focus on the content to be examined to heightened emphasis by students on academics--may be useful correctives to past problems of too much diversity in what is taught and too little student time and attention; others see them more negatively.

Positive Response to Testing from Teachers and Others

We found that Canadian teachers respond to the incentives offered them: many of them seek out the opportunities to be involved in provincial activities of setting curriculum standards and designing or grading tests of all kinds; they see these as valuable professional development efforts. Provincial authorities see them as building commitment to the results. Surveys and public opinion polls show that teachers and the Canadian public manifest general approval of the examination systems and believe that education has benefited.

Uncertainties Concerning Canadian National Test

Finally, we were interested to discover that the Canadian provinces have initiated a project to develop a national test. Because of the extensive province-level systems of standards and tests I have just described, the national project has encountered many objections. Agreement on the standards to be used has been elusive, and one province has decided that its disagreements are so fundamental that it will not take part at all. Extensive work has been done to define what is expected and how it should be measured. There is consensus that the purpose of the effort is monitoring and that there will thus be no stakes for students. Reporting is planned to extend no lower than the province level:

Canadian officials believe school or district-level monitoring by this national test would raise the stakes too high and compromise participation, as well as being much more costly owing to the larger samples needed. The present plan is for several provinces to work together to develop, on behalf of all the provinces involved, a new test to measure the standards emerging from the multiprovince conversations. However, many key matters remain unsettled, including disagreements within professional groups about the emphasis to be given different topics within a subject and the testing methods to be used, the level of difficulty of the test, and disagreements between educators and employers about the balance of academic and real-world skills to be tested.

Summary of Observations on Standards and Testing in Canada

In short, in Canada we found a coordinated set of standards, course specifications, and tests that are well-regarded by both educators and the public. Monitoring and accountability purposes are separated, and teachers are extensively involved in the activities of deciding what should be learned and of measuring the results. However, we could find no strong information on the effectiveness of Canada's system; it is implemented essentially at the provincial level; and efforts among the provinces to gain consensus on a plan for common standards and a national test have proven to be of great difficulty and uncertain feasibility.

CONCLUDING OBSERVATIONS

Let me conclude, Mr. Chairman, with some general observations that link the details of our work to broader questions of testing policy facing the Congress.

National Testing Design Must Flow From Purpose

First, is national testing feasible? Although our data show some skepticism on the part of officials and educators, they were reacting only to a general concept of expanded national testing. The Canadian experience suggests that the key determinant of feasibility may be deciding on the purpose to be achieved by testing. This is because most issues of technical quality (for example, validity and reliability) and cost must be addressed in a specific context of purpose. For example, if the purpose is monitoring, samples can be used that afford great flexibility in the type of test and large cost savings, even if expensive testing methods are used. If the purpose is accountability, such as certifying students, tests must have safeguards and other properties that will be expensive, including security; also, equitable exposure to the tested material is of critical importance to a fair use of the results. Maximizing one purpose may degrade another: the research shows that the higher the stakes of a test, the more effort individuals will put into assuring high scores quite apart from genuine learning, which in

turn makes the data less valid for monitoring. Our sense is that the debate over national tests has not yet distinguished clearly among the purposes to be served, nor has it drawn the appropriate conclusions concerning the technical difficulties involved in reconciling the conflicting requirements of a multipurpose test. We found the Canadian observations helpful in showing the feasibility of separate testing systems clearly specified to serve different purposes.

Finally, and again with respect to feasibility, our estimates suggest that students are not currently overtested and that the likely resource expenditures for various national test options are not exorbitant. However, these expenditures could vary considerably, depending on the purposes that are chosen for the test. At present, we have an open field of options before us, with none foreclosed. Yet it is not clear that we can achieve all purposes with one test.

The Desire for Rapid Development Must Not Constrain the Technical Quality of Measures

Second, will measurement be accurate? Both policy decisions, in general, and decisions affecting individual students and teachers should rest on sound data. Here, the key question is how we intend to test. For now, our hopes outstrip our capacity. As our respondents showed us, there is a yearning for better ways to test, so that we do full justice to students' learning, yet there is uncertainty over the state of the art in testing once we go beyond the familiar methods and a recognition of danger in the overeager use of unproven measures. We do not know whether the intense pressure first seen in 1990 and 1991 for the immediate implementation of a national test has abated, but we do know that high-quality innovative measurements, especially if adapted to many different regions (NCEST's clusters of states), will not be done quickly. Funding and governance arrangements need, therefore, to include careful monitoring of the technical aspects of the work, so that eagerness for rapid results does not supplant quality as the prime goal.

Standards Raise Many Tough Issues Worth Considerable Effort in Design and Implementation

Third, and last, where should we begin? Just the initial step of setting standards for student learning is quite difficult and it raises procedural, conceptual, and technical issues such as what roles should be played by educators and others in setting standards, what is needed by all or only some students, and how can such efforts guard against setting standards that are technically unmeasurable? Groups are at work in the United States on precisely this, some with federal support, for many different subjects, but the work has just begun and not all the

issues are on the table yet. And we must acknowledge that, in general, our schools do not now hew to high standards for rigorous academic work for all students. That is, the set of new content standards will pose considerable challenges for teachers and students, quite apart from the measurement problems we have just discussed. How will time be found to teach all the new material likely to be urged by each subject-matter group? What about schools that lack the instructional resources needed or teachers who lack the knowledge to be covered? And as implementation begins to affect measurement, what happens to test comparability if states and districts cannot handle all the material required by new standards and make different choices among the new requirements?

Given so much complexity, it may be wise to begin by emphasizing work on standards for the next several years, including some of the thornier issues of how they will come alive within the schools, while allowing the many promising state experiments with testing methods and formats to yield their results. In this way, we can learn much more about what works before we take too many major decisions about what and how to test at national levels. That would allow the debate over purpose to catch up as well, which is critical because practical choices will be difficult without a resolution of that debate and because the final answers to the questions of feasibility, cost and measurement accuracy also depend on purpose.

MATTERS FOR CONSIDERATION

We would emphasize two matters. First, because of the sizable knowledge base in current testing programs, because of the voices of opposition and uncertainty we heard from our survey respondents, and because of the successful Canadian experience with regard to teacher involvement, we believe it would be important for the Congress to consider specific ways to encourage the participation of teachers as well as state and local education administrators in further steps of developing standards and all aspects of increased testing (including development, administration, and scoring).

Second, we believe that the Congress should carefully consider how to ensure the technical quality of any tests in a national examination system. This is not only because technical quality was a frequently reported concern of our respondents but also because of the combined popularity and newness of large-scale performance testing. Popularity often results in time pressures and compressed schedules, whereas newness requires the development of valid, reliable tests and efficient and reliable scoring methods, all of which need trial, effort, and time. Seven states have performance tests now, and nine others told us they are 3 years away from such tests; creating a national system will be an effort of unprecedented scope and novelty and yet

enthusiastic prompting for immediate action seems likely. Money and time can always be saved at the expense of quality: for example, by doing less pilot testing, creating fewer test forms, shortening the test, or relaxing test security. In view of the lasting effects of incorrect decisions based on flawed test data, we urge explicit and proactive consideration to quality assurance in any national examination system implementation plan.

Mr. Chairman, this concludes my statement. I will be happy to answer any questions.

(973736, 740, 741)

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20884-6015

or visit:

Room 1000
700 4th St. NW (corner of 4th and G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (301) 258-4066.

United States
General Accounting Office
Washington, D.C. 20548

Official Business
Penalty for Private Use $300

First-Class Mail
Postage & Fees Paid
GAO
Permit No. G100